IDŐJÁRÁS

Calibration of wind speed ensemble forecasts for power generation

Sándor Baran* and **Ágnes Baran**

Faculty of Informatics, University of Debrecen
Kassai út 26, H-4028 Debrecen, Hungary

*Corresponding author E-mail: baran.sandor@inf.unideb.hu

Abstract— In the last decades, wind power became the second largest energy source in the EU covering 16% of its electricity demand. However, due to its volatility, accurate short range wind power predictions are required for successful integration of wind energy into the electrical grid. Accurate predictions of wind power require accurate hub height wind speed forecasts, where the state-of-the-art method is the probabilistic approach based on ensemble forecasts obtained from multiple runs of numerical weather prediction models. Nonetheless, ensemble forecasts are often uncalibrated and might also be biased, thus require some form of post-processing to improve their predictive performance. We propose a novel flexible machine learning approach for calibrating wind speed ensemble forecasts, which results in a truncated normal predictive distribution. In a case study based on 100m wind speed forecasts produced by the operational ensemble prediction system of the Hungarian Meteorological Service, the forecast skill of this method is compared with the predictive performance of three different ensemble model output statistics approaches and the raw ensemble forecasts. We show that compared with the raw ensemble, post-processing always improves the calibration of probabilistic and accuracy of point forecasts, and from the four competing methods, the novel machine learning based approach results in the best overall performance.

Key-words: ensemble calibration, ensemble model output statistics, multilayer perceptron, wind energy, wind speed

1. Introduction

The increasing challenges caused by consequences of air pollution and emission of greenhouse gases highlight the importance of transition of energy production towards renewable energy sources. Besides the classical hydro power, in the last decades, photovoltaic and wind energy fulfilled larger and larger part of energy demand. In 2020, the world set a new record by adding 93 GW of new wind turbines, so the total capacity of wind farms reached 744 GW covering 7% of the global electricity demand (*World Wind Energy Association,* 2021). In the EU (United Kingdom included), this proportion reached 16%, and the (world) record is held by Denmark, where wind accounted for 48% of the electricity consumed in 2020 (*Wind Europe,* 2021). However, wind energy poses serious challenges to traditional electricity markets, so accurate short range (between several minutes and a couple of days) prediction of wind power is of utmost importance for wind farm managers and electric grid operators.

Although the relation between wind speed and produced wind energy is nonlinear and might also be nonstationary, more reliable wind speed forecasts obviously result in more reliable predictions of produced electricity. Wind speed forecasts, similar to other meteorological variables, are based on numerical weather prediction (NWP) models describing atmospheric processes via systems of partial differential equations. The state of the art approach is to run an NWP model several times with different initial conditions which results in an ensemble of forecasts (*Bauer et al.,* 2015). Ensemble forecasts enable estimation of situation dependent probability distributions of future weather variables, which opens the door for probabilistic weather forecasting (*Gneiting* and *Raftery,* 2005), where besides getting a point forecast, the forecast uncertainty is also assessed.

Recently, all major weather centres operate their own ensemble prediction system (EPS), e.g., the 35-member Prévision d'Ensemble ARPEGE[1] (PEARP) EPS of Météo France (*Descamps et al.,* 2015) or the 11-member Applications of Research to Operations at Mesoscale EPS (AROME-EPS; *Jávorné Radnóczi et al.,* 2020) of the Hungarian Meteorological Service (HMS), whereas the largest ensemble size corresponds to the 51-member EPS of the European Centre for Medium-Range Weather Forecasts (*Buizza et al.,* 1998). Nowadays ensemble weather forecasts are also popular inputs to probabilistic forecasts of renewable energy (*Pinson* and *Messner,* 2018).

However, ensemble forecasts often appear to be uncalibrated and/or biased, this feature has been observed in several operational ensembles (see, e.g., *Buizza et al.,* 2005). A possible solution is the use of some form of statistical post-processing (*Buizza,* 2018), where nonparametric methods usually capture predictive distributions via estimating their quantiles (see, e.g., *Friederichs* and *Hense,* 2007; *Bremnes,* 2019), whereas parametric post-processing approaches

---
[1] Action de Recherche Petite Echelle Grande Echelle (i.e. Research Project on Small and Large Scales)

provide full predictive distributions of the future weather quantities (see, e.g., *Gneiting et al.,* 2005; *Raftery et al.,* 2005). Recently, machine learning based methods also gain more and more popularity (see, e.g., *Rasp* and *Lerch,* 2018; *Taillardat* and *Mestre,* 2020); for a detailed overview of statistical calibration techniques we refer to *Wilks* (2018) or *Vannitsem et al.* (2021).

Here we focus on a weather quantity important in energy production and investigate statistical post-processing of ensemble forecasts of wind speed measured at hub height (100m). In this context, *Tailor et al.* (2009) proposes kernel dressing with Gaussian kernel left truncated at zero (TN; truncated normal), while *Messner et al.* (2013) considers forecasts based on inverse power curves and applies a censored normal predictive distribution. However, any post-processing method appropriate for wind speed can be applied, and we concentrate on the ensemble model output statistic (EMOS; *Gneiting et al.,* 2005) approach, where the predictive distribution is a single parametric probability law with parameters depending on the ensemble forecasts via appropriate link functions. To account for the non-negativity and right skew of wind speed, *Thorarinsdottir* and *Gneiting* (2010) proposes a TN, *Baran* and *Lerch* (2015) a log-normal (LN), whereas *Baran et al.* (2021) a truncated generalized extreme value (TGEV) predictive distribution, and several methods for combining these probabilistic forecasts have also been developed (see, e.g., *Lerch* and *Thorarinsdottir,* 2013; *Baran* and *Lerch,* 2016, 2018).

In the present paper we test the forecast skill of TN, LN, and TGEV EMOS approaches on AROME-EPS forecasts of hub height wind speed. We also introduce a novel model with TN predictive distribution, where using the ideas of *Rasp* and *Lerch* (2018) and *Ghazvinian et al.* (2021), location and scale parameters of the TN law are connected to the ensemble members via a multilayer perceptron neural network (MLP; *Goodfellow et al.,* 2016). Compared with the case of fixed link functions, this latter approach allows more flexibility in modeling and straightforward inclusion of new covariates as well. Note that TN, LN, and TGEV EMOS approaches and some of their combinations have already been successfully applied for calibration of surface wind speed forecasts of the 11-member Aire Limitée Adaptation dynamique Développement International-Hungary Ensemble Prediction System of the HMS (*Horányi et al.,* 2006), see, e.g., *Baran et al.* (2014).

The paper is organized as follows. In Section 2, the detailed description of the AROME-EPS is given, while in Section 3, the applied post-processing methods and considered verification tools are reviewed. The results of our case study is presented in Section 4 followed by a concluding discussion in Section 5.

2. Data

The 11-member AROME-EPS of the HMS covers the Transcarpathian Basin with a horizontal resolution of 2.5 km (*Jávorné Radnóczi et al., 2020*). It consists of a control member and 10 ensemble members obtained from perturbed initial conditions. The dataset at hand contains ensemble forecasts of wind speed (m/s) at hub height (100m) together with the corresponding validation observations for three wind farms in the northwestern part of Hungary (Ács, Jánossomorja, and Pápakovácsi) for the period May 7, 2020 to March 28, 2021. All forecasts are initialized at 0000 UTC with a temporal resolution of 15 minutes and maximal forecast horizon of 48 h resulting in a total of 192 forecast lead times.

3. Post-processing methods and verification tools

Non-homogeneous regression or EMOS is one of the most popular parametric post-processing approaches, probably due to its computational efficiency and excellent performance for a wide range of weather variables. EMOS models for different weather quantities differ in the parametric family specifying the predictive distribution; however, most of the existing EMOS models are implemented in the ensembleMOS package of R (*Yuen et al., 2018*).

In the following sections let $f_1, f_2, \ldots, f_{11}$ denote the 11-member AROME-EPS hub height wind speed forecast for a given location, time, and lead time, where $f_1 = f_{\text{CTRL}}$ is the control forecast, while $f_2, f_3, \ldots, f_{11}$ correspond to the 10 statistically indistinguishable (and thus exchangeable) ensemble members $f_{\text{ENS},1}, f_{\text{ENS},2}, \ldots, f_{\text{ENS},10}$ generated using random perturbations. Further, let $\overline{f}$ denote the ensemble mean, $\overline{f}_{\text{ENS}}$ denote the mean of the 10 exchangeable members, and S^2 and MD denote the ensemble variance and ensemble mean absolute difference, respectively, defined as

$$S^2 := \frac{1}{10} \sum_{k=1}^{11} (f_k - \overline{f})^2 \qquad \text{and} \qquad \text{MD} := \frac{1}{11^2} \sum_{k=1}^{11} \sum_{\ell=1}^{11} |f_k - f_\ell|.$$

3.1. Truncated normal EMOS model

Let $\mathcal{N}_0(\mu, \sigma^2)$ denote the TN distribution with location μ, scale $\sigma > 0$, and lower truncation at 0, having probability density function (PDF)

$$g(x|\mu, \sigma) := \frac{1}{\sigma} \varphi((x - \mu)/\sigma)/\Phi(\mu/\sigma), \qquad \text{if } x \geq 0,$$

and $g(x|\mu, \sigma) := 0$, otherwise, where φ is the PDF; while Φ denotes the cumulative distribution function (CDF) of a standard normal distribution. The proposed TN EMOS predictive distribution for hub height wind speed based on the AROME-EPS ensemble forecast is

$$\mathcal{N}_0(a_0 + a_{\mathrm{CTRL}}^2 f_{\mathrm{CTRL}} + a_{\mathrm{ENS}}^2 \overline{f}_{\mathrm{ENS}}, b_0^2 + b_1^2 \mathrm{MD}), \tag{1}$$

where $a_0, a_{\mathrm{CTRL}}, a_{\mathrm{ENS}}, b_0, b_1 \in \mathbb{R}$. The same model is applied by *Hemri et al.* (2014) to model square root of 10m wind speed, and the suggested method is a slight modification of the TN EMOS approach of *Thorarinsdottir* and *Gneiting* (2010), where the square of the scale parameter is an affine function of the ensemble variance, that is $\sigma^2 = b_0^2 + b_1^2 S^2$. Exploratory tests with the dataset at hand show that neither modelling the square root of the data, nor linking location to the ensemble variance result in better forecast skill than the use of Eq. (1).

3.2. Log-normal EMOS model

As an alternative to the TN EMOS approach, we consider the EMOS model of *Baran* and *Lerch* (2015), where the mean m and variance v of the LN predictive distribution are affine functions of the ensemble members and the ensemble variance, respectively, that is

$$m = \alpha_0 + \alpha_{\mathrm{CTRL}}^2 f_{\mathrm{CTRL}} + \alpha_{\mathrm{ENS}}^2 \overline{f}_{\mathrm{ENS}} \quad \text{and} \quad v = \beta_0^2 + \beta_1^2 S^2,$$

where $\alpha_0, \alpha_{\mathrm{CTRL}}, \alpha_{\mathrm{ENS}}, \beta_0, \beta_1 \in \mathbb{R}$. The heavier upper tail of the LN distribution allows a better fit to high wind speed values.

3.3. Truncated generalized extreme value EMOS model

Another possible solution to address reliability of probabilistic forecasts for high wind speed is the use of the GEV EMOS approach proposed by *Lerch* and *Thorarinsdottir* (2013). The GEV distribution $\mathcal{GEV}(\mu, \sigma, \xi)$ with location μ, scale $\sigma > 0$, and shape ξ is defined by CDF

$$G(x|\mu, \sigma, \xi) := \begin{cases} \exp(-[1 + \xi(\frac{x-\mu}{\sigma})]^{-1/\xi}), & \text{if } \xi \neq 0; \\ \exp(-\exp(-\frac{x-\mu}{\sigma})), & \text{if } \xi = 0, \end{cases}$$

for $1 + \xi(\frac{x-\mu}{\sigma}) > 0$ and $G(x|\mu, \sigma, \xi) := 0$, otherwise. However, as demonstrated by *Lerch* and *Thorarinsdottir* (2013) and *Baran* and *Lerch* (2015), the GEV EMOS model might assign positive predicted probability to negative wind

speed. To correct this deficiency, *Baran et al.* (2021) proposed to truncate the GEV distribution from below at zero and considered a TGEV predictive distribution $\mathcal{TGEV}(\mu, \sigma, \xi)$ with location μ, scale $\sigma > 0$, and shape ξ defined by CDF

$$G_0(x|\mu,\sigma,\xi) := \begin{cases} \frac{G(x|\mu,\sigma,\xi) - G(0|\mu,\sigma,\xi)}{1 - G(0|\mu,\sigma,\xi)}, & \text{if } G(0|\mu,\sigma,\xi) < 1; \\ 1, & \text{if } G(0|\mu,\sigma,\xi) = 1, \end{cases}$$

for $x \geq 0$, and $G(x|\mu,\sigma,\xi) := 0$, otherwise.

For the 11-member AROME-EPS, location and scale parameters of the TGEV EMOS model are

$$\mu = \gamma_0 + \gamma_{\text{CTRL}} f_{\text{CTRL}} + \gamma_{\text{ENS}} \overline{f}_{\text{ENS}} \qquad \text{and} \qquad \sigma = \sigma_0^2 + \sigma_1^2 \overline{f},$$

with $\gamma_0, \gamma_{\text{CTRL}}, \gamma_{\text{ENS}}, \sigma_0, \sigma_1 \in \mathbb{R}$, while the shape parameter ξ does not depend on the ensemble members. In order to ensure a finite mean and positive skewness, the shape is kept in the interval $]-0.278, 1/3[$.

3.4. Parameter estimation

Parameter estimation in the TN, LN, and TGEV EMOS models described in Sections 3.1 – 3.3 is based on the optimum score principle of *Gneiting* and *Raftery* (2007). The estimates are obtained as minimizers of the mean value of a proper scoring rule over an appropriate training dataset. Here we consider one of the most popular proper scores in atmospheric sciences, namely the continuous ranked probability score (CRPS; *Wilks,* 2019, Section 9.5.1). Given a (predictive) CDF F and a real value (observation) x, the CRPS is defined as

$$\text{CRPS}(F, x) := \int_{-\infty}^{\infty} [F(y) - \mathbb{I}_{\{y \geq x\}}]^2 dy = \text{E}|X - x| - \frac{1}{2}\text{E}|X - X'|, \quad (2)$$

where $\mathbb{I}_H$ denotes the indicator function of a set H, while X and X' are independent random variables distributed according to F and having a finite first moment. CRPS is a negatively oriented score, that is the smaller the better, and the right-hand side of Eq. (2) implies that it can be expressed in the same units as the observation. Note that the CRPS for TN, LN, and TGEV distributions can be expressed in closed form (see *Thorarinsdottir* and *Gneiting* (2010), *Baran* and *Lerch* (2015), and *Baran et al.* (2021), respectively), which allows an efficient optimization procedure.

A crucial issue in statistical calibration is the selection of training data. Here the different forecast horizons are treated separately, and we use rolling training periods, which is a standard approach in EMOS modeling. In this training scheme, parameters for a given lead time are estimated with the help of

corresponding forecast–observation pairs from the preceding n calendar days. Further, both regional (or global) and local EMOS models are investigated. In the regional approach, all data from the training period are considered together, providing a single set of EMOS parameters for all three wind farms. In contrast, local estimation results in different parameter estimates for different wind farms by using only data of the given location. In general, local models outperform their regional counterparts (see, e.g., *Thorarinsdottir* and *Gneiting,* 2010), provided the training period is long enough to avoid numerical stability issues (*Lerch* and *Baran,* 2017).

3.5. *Machine learning based approach to wind speed modeling*

As mentioned in the Introduction, based on works of *Rasp* and *Lerch* (2018) and *Ghazvinian et al.* (2021), we applied a machine learning approach to estimate the parameters of the predictive distribution in a TN model. In this case, instead of looking for the parameters $a_0, a_{\text{CTRL}}, a_{\text{ENS}}, b_0, b_1$ in Eq. (1), location and scale are estimated directly, without assuming that they depend on the ensemble in a prescribed way. Practically this means, that some features derived from the ensemble (e.g., the control member, or the ensemble standard deviation) are used as inputs of a multilayer perceptron (MLP), while the trained network provides a two-dimensional vector corresponding to the location and scale parameters. Similar to the previous models, the network is trained by minimizing the mean CRPS over the training data.

In an MLP. some hidden layers connect the input layer and the output one, the number of layers and the number of neurons in the different hidden layers are tuning parameters of the network. Starting from the first hidden layer, each neuron of the given layer computes a weighted sum of the values provided by the neurons in the previous layer, adds a bias, and via a so-called transfer function, applies a transformation to the result.

In the present work we train an MLP with one hidden layer containing 25 neurons, the applied transfer functions are the exponential linear unit (ELU; see e.g., *Ghazvinian et al.,* 2021) function in the hidden layer, and the linear function in the output layer. After some experiments, in the final training we decided to use the control forecast, the mean of the exchangeable ensemble members, and the standard deviation of the 11 members as input features of the network. Based on *Ghazvinian et al.* (2021), to ensure the positivity of the location and scale parameters, their estimates are given by $\exp(\theta_1)$ and $\exp(\theta_2)$, where θ_1 and θ_2 are the values provided by the two neurons of the output layer.

By the training of a network, the number of the training samples is always a critical point: a relatively small training set can easily result in overfitting, which means a weak performance on the test set. In order to avoid this problem, we apply a regional estimation, moreover, we do not handle the different lead times separately; for a given training period we train only two networks, one for the 0–

24h forecasts, another for the 24–48h forecasts. We made a trial to take into account the lead time in the training by extending the features with a fourth one, containing the ranks of the lead times; however, this modification did not improve the predictive performance of the network. The lack of significance of the forecast horizon might be explained by the diurnal cycle in the ensemble standard deviation, which indicates a direct relation between forecast uncertainty and lead time.

3.6. Verification tools

As argued by *Gneiting et al.* (2007), the aim of probabilistic forecasting is to maximize the sharpness of the predictive distribution subject to calibration. The former refers to the concentration of the predictive distribution, whereas the latter means a statistical consistency between the validating observation and the corresponding predictive distribution. These goals can be addressed simultaneously using proper scoring rules quantifying the forecast skill by numerical values assigned to pairs of probabilistic forecasts and validating observations. In the case study of Section 4, for a given lead time, competing forecasts in terms of probability distribution are compared with the help of the mean CRPS over all forecast cases in the verification data. The improvement in terms of CRPS of a probabilistic forecast F with respect to a reference forecast F_{ref} can be assessed with the continuous ranked probability skill score (CRPSS; see, e.g., *Gneiting* and *Raftery*, 2007) defined as

$$\text{CRPSS} := 1 - \frac{\overline{\text{CRPS}_F}}{\overline{\text{CRPS}_{F_{\text{ref}}}}},$$

where $\overline{\text{CRPS}_F}$ and $\overline{\text{CRPS}_{F_{\text{ref}}}}$ denote the mean score values corresponding to forecasts F and F_{ref}, respectively. Here larger values indicate better forecast skill compared to the reference method.

Calibration and sharpness can also be quantified by the coverage and average width of the $(1 - \alpha)100\,\%$, $\alpha \in\,]0,1[$, central prediction interval, where calibration is defined as the proportion of validating observations located between the lower and upper $\alpha/2$ quantiles of the predictive distribution. For a well calibrated forecast, this value should be around $(1 - \alpha)100\,\%$, and in order to provide a fair comparison with the 11-member AROME-EPS, α should be chosen to match the nominal coverage of $83.33\,\%$ ($10/12 \times 100\,\%$) of the raw ensemble.

Simple graphical tools for assessing calibration of probabilistic forecasts are the verification rank histogram of ensemble predictions and its continuous counterpart, the probability integral transform (PIT) histogram. Verification rank is defined as the rank of the verifying observation with respect to the corresponding ensemble forecast (*Wilks,* 2019, Section 9.7.1), whereas PIT is the value of the predictive CDF evaluated at the observation (*Wilks,* 2019, Section 9.5.4). For a

properly calibrated ensemble, all ranks should be equally likely, while calibrated predictive distributions result in standard uniform PIT values.

Finally, the accuracy of point forecasts, such as median and mean, is quantified with the help of mean absolute errors (MAEs) and root mean squared errors (RMSEs), respectively.

4. Results

We start our analysis by determining the appropriate training-period length for our post-processing approaches. We consider a fixed verification period from July 8, 2020 to March 28, 2021 (264 calendar days) and compare the forecast skill of both local and regional TN EMOS models estimated using $20, 21, \ldots, 60$ day rolling training-periods. *Fig. 1* shows the mean CRPS taken over all forecast cases and lead times and the MAE of median forecasts as functions of the training-period length. Both plots clearly demonstrate that for longer training periods, the local TN EMOS is more skillful than the regional one. CRPS and MAE of the latter stabilize after day 51, while the corresponding scores of the local TN EMOS also seem to level off there. Hence, for TN EMOS modeling, a 51-day training-period seems to be a reasonable choice, and the same training-period length is applied for LN and GEV EMOS models as well. A detailed data analysis confirmed that this length is also appropriate for the machine learning approach of Section 3.5 (TN MLP), this choice of training data leaves a total of 273 calendar days (period June 29, 2020 – March 28, 2021) for model verification. Further, as in general, local versions of the tested EMOS approaches slightly outperform the regional ones, thus, in what follows, only the scores of the local models are reported.
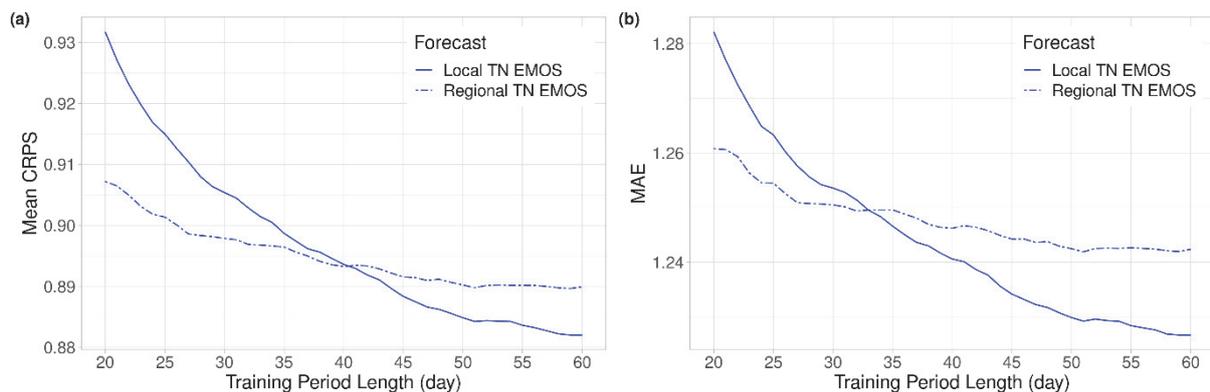

Fig. 1. Mean CRPS of probabilistic (a) and MAE of median (b) forecasts for local and regional TN EMOS models as functions of training-period length.

Fig. 2a shows the mean CRPS of post-processed and raw ensemble forecasts as functions of the lead time, whereas in *Fig. 2b*, the corresponding CRPSS values with respect to the raw ensemble are plotted. In general, all post-processing approaches outperform the raw ensemble for all lead times, but the advantage of post-processing decreases with the increase of the forecast horizon. The best overall CRPSS taken over all lead times and forecast cases belongs to the TN MLP model (0.111), followed by the local TN EMOS method (0.103); however, there are certain forecast horizons (especially around 20h and 23h), where the latter exhibits slightly better predictive performance. For the TGEV and LN EMOS approaches, these overall CRPSS values are 0.091 and 0.095, respectively.
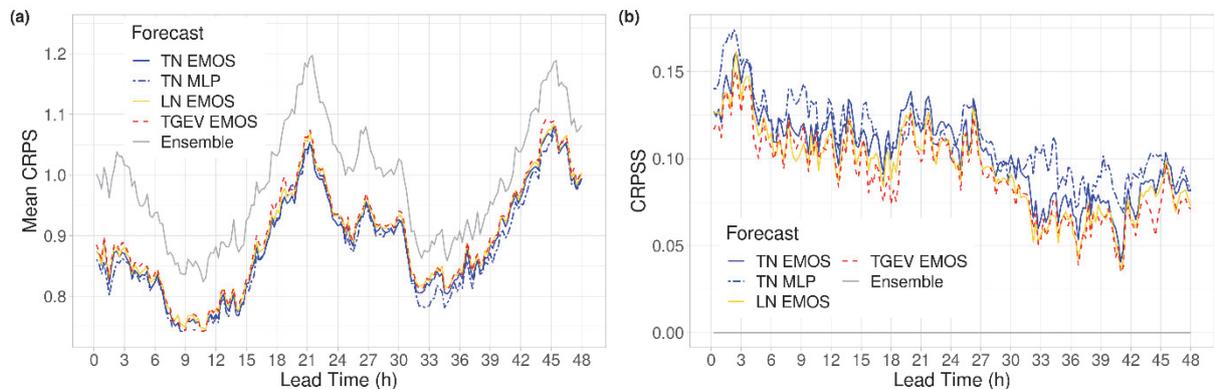

Fig. 2. Mean CRPS of post-processed and raw ensemble forecasts of wind speed (a) and CRPSS with respect to the raw ensemble (b) as functions of lead time.

The improved calibration of post-processed forecasts can also be observed in *Fig. 3a* showing the coverage of the nominal 83.33% central prediction intervals for different lead times. The coverage of the AROME-EPS ranges from 50% to 70%, and in general, increases with the increase of the lead time, whereas all post-processed forecasts for all lead times result in coverage values that are rather close to the nominal level. In particular, there is no visible systematic difference in the coverage values of the three investigated EMOS models, whereas the TN MLP approach seems to exhibit some kind of diurnal cycle. However, as depicted in *Fig. 3b*, the cost of the better calibration should be paid in the deterioration of the sharpness. The raw ensemble produces far the narrowest central predictive intervals, there is no difference in sharpness between the competing EMOS models, whereas the diurnal cycle in sharpness of the TN MLP is completely in line with the corresponding coverage. Note that similar diurnal cycles can be observed in the ensemble standard deviation and ensemble mean difference as well.
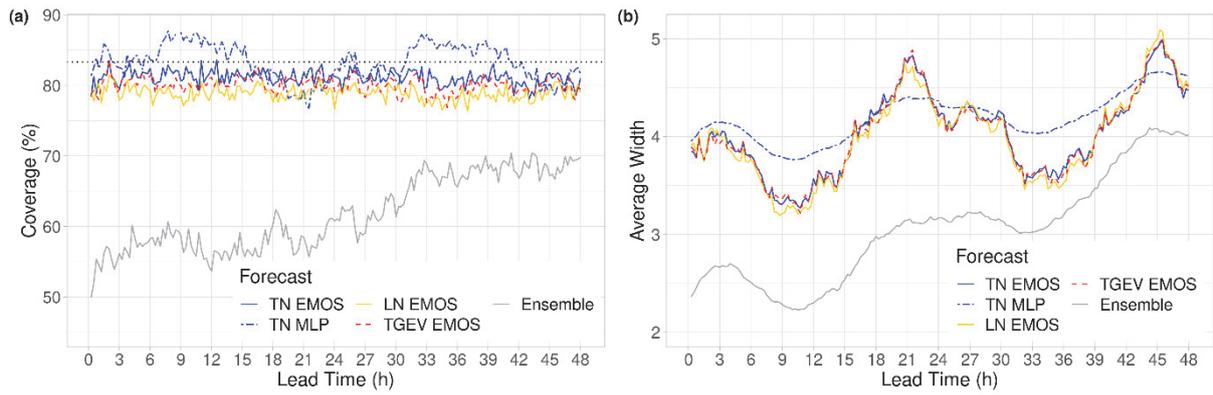
Fig. 3. Coverage (a) and average width (b) of the nominal 83.33 % central prediction intervals of post-processed and raw forecasts as functions of lead time.

While statistical post-processing substantially improves the calibration of probabilistic forecasts, it does not really effect the accuracy of point predictions. In *Fig. 4a*, the difference in MAE of the median forecasts of the various calibration methods from the MAE of the raw ensemble are plotted as functions of the lead time. Similar to the mean CRPS, models with TN predictive distribution show the best performance for all lead times; however, even the largest difference in MAE is less than 0.1 m/s. The same behavior can be observed in *Fig. 4b* displaying the difference in RMSE of the mean forecasts. This can indicate that the raw AROME-EPS forecasts are already unbiased and indeed, the mean biases of the ensemble mean and median taken over all forecast cases of the whole available period May 8, 2020 to March 28, 2021 and all lead times are just 0.136 m/s and 0.122 m/s, respectively, while the overall MAE equals 1.285 m/s and the overall RMSE is 1.669 m/s.
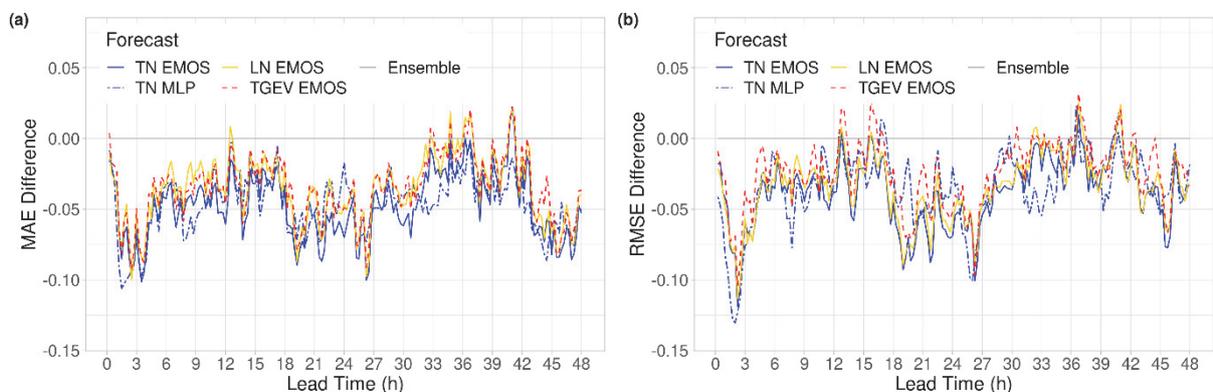

Fig. 4. Difference in MAE of the median forecasts (a) and in RMSE of the mean forecasts (b) from the raw ensemble as functions of lead time.

Finally, *Fig. 5* shows the verification rank histograms of raw and PIT histograms of post-processed forecasts for four different lead time intervals. The U-shaped verification rank histograms clearly indicate the underdispersive character of the raw ensemble; however, the dispersion improves with the forecast lead time. This behavior is completely in line with the increasing coverage and high sharpness of the raw forecasts (see *Fig. 3*). Further, the depicted rank histograms are rather symmetric, which is consistent with the small overall MAE and RMSE and illustrates the lack of bias in the raw ensemble. All post-processing approaches substantially improve calibration; models based on TN predictive distributions result in almost flat PIT histograms, whereas the histograms of TGEV and LN EMOS approaches indicate slight biases. Kolmogorov–Smirnov (KS) test rejects the uniformity of the PIT for all models; however, based on the values of the KS test statistic, one can provide a clear ranking of the methods. PIT values of the TN MLP approach fit best the uniform distribution, followed by the TN, TGEV, and LN EMOS models, which order nicely reflects the shapes of the corresponding histograms of *Fig. 5*.
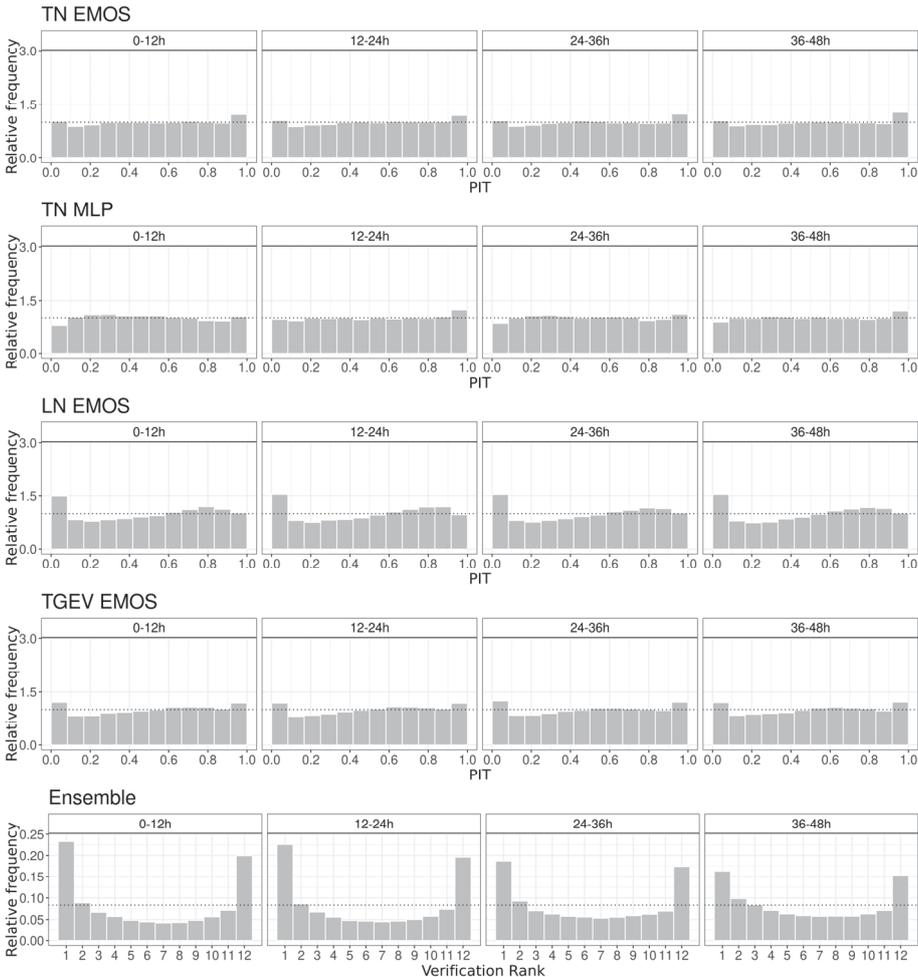

Fig. 5. PIT histograms of post-processed and verification rank histograms of raw ensemble forecasts of wind speed for the lead times 0-12h, 12-24h, 24-36h, and 36-48h.

Based on the above results, one can conclude that models with TN predictive distributions provide the best forecast skill, and the machine learning based TN MLP approach outperforms the TN EMOS model. Hence, one might be interested in the dissimilarities of the corresponding predictive distributions. According to *Fig. 6a*, there is no fundamental difference in the location, and the station-wise time series plots of this parameter also provide matching curves (not shown). Thus, the linear model of the location given in Eq. (1) seems to be optimal. A completely different picture can be observed in *Fig. 6b*, showing the mean of the scales of the TN predictive distributions as function of lead time. The diurnal cycle for TN MLP is far less pronounced than for the TN EMOS, and the corresponding time series (not shown) exhibit completely different behavior, too. Hence, the superior performance of the TN MLP approach is due to the more general modeling of the scale of the TN predictive distribution.
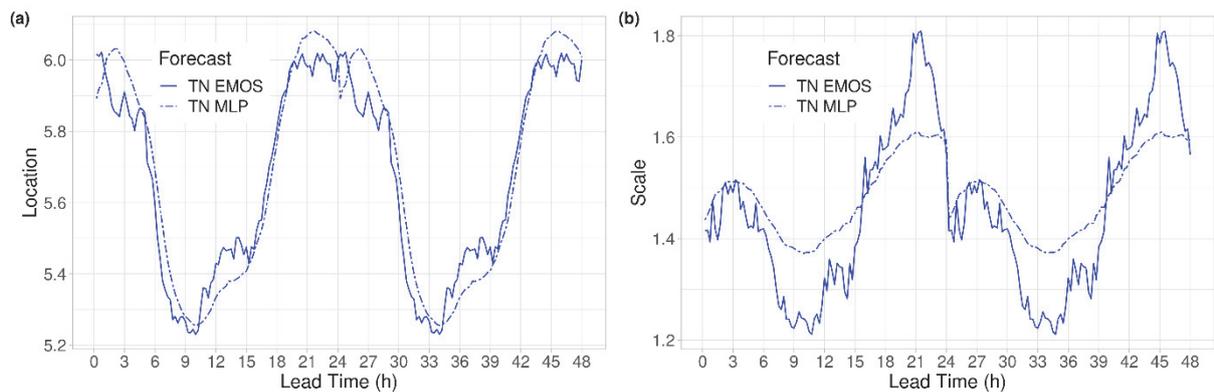

Fig. 6. Mean of the location (a) and scale (b) of the truncated normal predictive distributions of TN EMOS and TN MLP models as functions of lead time.

5. Conclusions

We investigate post-processing of ensemble forecasts of 100m wind speed, as this variable is of crucial interest in wind energy production. Three different EMOS models based on truncated normal, log-normal, and truncated generalized extreme value distributions are considered, and we also propose a novel method where the probabilistic forecasts are obtained in the form of a truncated normal predictive distribution with parameters linked to the ensemble via a multilayer perceptron neural network. The forecasts skill of the competing calibration methods is tested on the 11-member AROME-EPS hub height wind

speed ensemble forecasts of the HMS for three wind farms in Hungary and verified against observations provided by the wind farm operators. Only short-term predictions are considered with forecast horizons ranging up to 48 h with a temporal resolution of 15 minutes. Using the raw ensemble as reference, we compare the mean CRPS of probabilistic, MAE of median, and RMSE of mean forecasts, and the coverage of central prediction intervals corresponding to the nominal 83.33% coverage. We also study the shapes of the PIT histograms of the calibrated forecasts for different lead times and compare with the corresponding verification rank histograms of the raw ensemble. Based on our case study we can conclude, that compared with the raw ensemble, post-processing always improves the calibration of probabilistic and accuracy of point forecasts. From the four competing methods, the novel machine learning based TN MLP approach exhibits the best overall performance; moreover, in contrast to the investigated EMOS models, it provides a single universal model for several forecast horizons. The superior performance of the TN MLP model is explained by its ability to represent more complex nonlinear relations between the ensemble forecasts and the parameters of the TN predictive distribution, and our results are consistent with the findings of *Rasp* and *Lerch* (2018) and *Ghazvinian et al.* (2021).

The present work highlights several directions of potential future research. From the one hand, one might consider the machine learning approach to parameter estimation in the case of other predictive distribution families such as the LN and TGEV investigated here. From the other hand, a neural network allows a very flexible choice of input features, providing a simple and straightforward opportunity of involving predictions of other weather variables in wind speed modeling.

References

Baran, S., Horányi, A. and *Nemoda, D.,* 2014: Comparison of the BMA and EMOS statistical methods in calibrating temperature and wind speed forecast ensembles. *Időjárás 118*, 217–241.

Baran, S. and *Lerch, S.,* 2015: Log-normal distribution based Ensemble Model Output Statistics models for probabilistic wind speed forecasting. *Q. J. R. Meteorol. Soc. 141*, 2289–2299. https://doi.org/10.1002/qj.2521

Baran, S. and *Lerch, S.,* 2016: Mixture EMOS model for calibrating ensemble forecasts of wind speed. *Environmetrics 27*, 116–130. https://doi.org/10.1002/env.2380

Baran, S. and *Lerch, S.,* 2018: Combining predictive distributions for statistical post-processing of ensemble forecasts. *Int. J. Forecast. 34*, 477–496. https://doi.org/10.1016/j.ijforecast.2018.01.005

Baran, S., Szokol, P., and *Szabó, M.,* 2021: Truncated generalized extreme value distribution based EMOS model for calibration of wind speed ensemble forecasts. *Environmetrics 32,* e2678. https://doi.org/10.1002/env.2678

Bauer, P., Thorpe, A. and *Brunet, G.,* 2015: The quiet revolution of numerical weather prediction. *Nature 525*, 47–55. https://doi.org/10.1038/nature14956

Bremnes, J.B., 2019: Constrained quantile regression splines for ensemble postprocessing. *Mon. Weather Rev. 147*, 1769–1780. https://doi.org/10.1175/MWR-D-18-0420.1

Buizza, R., 2018: Ensemble forecasting and the need for calibration. In (eds. *Vannitsem, S., Wilks, D. S., Messner, J.W.*), Statistical Postprocessing of Ensemble Forecasts, Elsevier, Amsterdam, 15–48. https://doi.org/10.1016/B978-0-12-812372-0.00002-9

Buizza, R., Houtekamer, P. L., Toth, Z., Pellerin, G., Wei, M., and *Zhu, Y.,* 2005: A comparison of the ECMWF, MSC, and NCEP global ensemble prediction systems. *Mon. Weather Rev. 133*, 1076–1097. https://doi.org/10.1175/MWR2905.1

Buizza, R., Petroliagis, T., Palmer, T., Barkmeijer, J., Hamrud, M., Hollingsworth, A., Simmons, A., and *Wedi, N.,* 1998: Impact of model resolution and ensemble size on the performance of an Ensemble Prediction System. *Q. J. R. Meteorol. Soc. 124*, 1935–1960. https://doi.org/10.1002/qj.49712455008

Descamps, L., Labadie, C., Joly, A., Bazile, E., Arbogast, P., and *Cébron, P.,* 2015: PEARP, the Météo-France short-range ensemble prediction system. *Q. J. R. Meteorol. Soc. 141*, 1671–1685. https://doi.org/10.1002/qj.2469

Friederichs, P. and Hense, A., 2007: Statistical downscaling of extreme precipitation events using censored quantile regression. *Mon. Weather Rev. 135*, 2365–2378. https://doi.org/10.1175/MWR3403.1

Ghazvinian, M., Zhang, Y., Seo, D-J., He, M., and *Fernando, N.,* 2021: A novel hybrid artificial neural network - parametric scheme for postprocessing medium-range precipitation forecasts. *Adv. Water Resour. 151*, paper 103907. https://doi.org/10.1016/j.advwatres.2021.103907

Gneiting, T., Balabdaoui, F., and *Raftery, A.E.,* 2007: Probabilistic forecasts, calibration and sharpness. *J. R. Stat. Soc. B 69*, 243–268. https://doi.org/10.1111/j.1467-9868.2007.00587.x

Gneiting, T. and *Raftery, A.E.,* 2005: Weather forecasting with ensemble methods. *Science 310*, 248–249. https://doi.org/10.1126/science.1115255

Gneiting, T. and *Raftery, A.E.,* 2007: Strictly proper scoring rules, prediction and estimation. *J. Amer. Statist. Assoc. 102*, 359–378. https://doi.org/10.1198/016214506000001437

Gneiting, T., Raftery, A.E., Westveld, A.H., and *Goldman, T.,* 2005: Calibrated probabilistic forecasting using ensemble model output statistics and minimum CRPS estimation. *Mon. Weather Rev. 133*, 1098–1118. https://doi.org/10.1175/MWR2904.1

Goodfellow, I., Bengio, Y., and *Courville, A.,* 2016: Deep Learning. MIT Press, Cambridge.

Hemri, S., Scheuerer, M., Pappenberger, F., Bogner, K., and *Haiden, T.,* 2014: Trends in the predictive performance of raw ensemble weather forecasts. *Geophys. Res. Lett. 41*, 9197–9205. https://doi.org/10.1002/2014GL062472

Horányi, A., Kertész, S., Kullmann, L., and *Radnóti, G.,* 2006: The ARPEGE/ALADIN mesoscale numerical modeling system and its application at the Hungarian Meteorological Service. *Időjárás 110*, 203–227.

Jávorné Radnóczi, K., Várkonyi, A., and *Szépszó, G.,* 2020: On the way towards the AROME nowcasting system in Hungary. *ALADIN-HIRLAM Newsletter 14*, 65–69.

Lerch, S. and *Baran, S.,* 2017: Similarity-based semi-local estimation of EMOS models. *J. R. Stat. Soc. C 66,* 29–51. https://doi.org/10.1111/rssc.12153

Lerch, S. and *Thorarinsdottir, T.L.,* 2013: Comparison of non-homogeneous regression models for probabilistic wind speed forecasting. *Tellus A 65*, paper 21206. https://doi.org/10.3402/tellusa.v65i0.21206

Messner, J.W., Zeileis, A., Bröcker, J., and *Mayr, G.J.,* 2013: Probabilistic wind power forecasts with an inverse power curve transformation and censored regression. *Wind Energy 17*, 1753–1766. https://doi.org/10.1002/we.1666

Pinson, P. and *Messner, J.W.,* 2018: Application of Postprocessing for Renewable Energy. In (eds. *Vannitsem, S., Wilks, D.S., Messner, J.W.*) Statistical Postprocessing of Ensemble Forecasts, Elsevier, 241–266. https://doi.org/10.1016/B978-0-12-812372-0.00009-1

Raftery, A. E., Gneiting, T., Balabdaoui, F., and *Polakowski, M.,* 2005: Using Bayesian model averaging to calibrate forecast ensembles. *Mon. Weather Rev. 133*, 1155–1174. https://doi.org/10.1175/MWR2906.1

Rasp, S. and Lerch, S. (2018) Neural networks for postprocessing ensemble weather forecasts. *Mon. Weather Rev. 146*, 3885–3900. https://doi.org/10.1175/MWR-D-18-0187.1

Taillardat, M. and *Mestre, O.,* 2020: From research to applications – examples of operational ensemble post-processing in France using machine learning. *Nonlin. Proc. Geophys. 27*, 329–347. https://doi.org/10.5194/npg-27-329-2020

Taylor, J., McSharry, P. and *Buizza, R.,* 2009: Wind power density forecasting using ensemble predictions and time series models. *IEEE Trans. on Energy Convers. 24*, 775–782. https://doi.org/10.1109/TEC.2009.2025431

Thorarinsdottir, T. L. and *Gneiting, T.,* 2010: Probabilistic forecasts of wind speed: Ensemble model output statistics by using heteroscedastic censored regression. *J. R. Stat. Soc. A 173*, 371–388. https://doi.org/10.1111/j.1467-985X.2009.00616.x

Vannitsem, S., Bremnes, J. B., Demaeyer, J., Evans, G. R., Flowerdew, J., Hemri, S., Lerch, S., Roberts, N., Theis, S., Atencia, A., Ben Boualègue, Z., Bhend, J., Dabernig, M., De Cruz, L., Hieta, L., Mestre, O., Moret, L., Odak Plenkovič, I., Schmeits, M., Taillardat, M., Van den Bergh, J., Van Schaeybroeck, B., Whan, K., and *Ylhaisi, J.,* 2021: Statistical postprocessing for weather forecasts – review, challenges and avenues in a big data world. *Bull. Amer. Meteorol. Soc. 102*, E681–E699. https://doi.org/10.1175/BAMS-D-19-0308.1

Wilks, D.S., 2018: Univariate ensemble forecasting. In (eds. *Vannitsem, S., Wilks, D.S., Messner, J.W.*), Statistical Postprocessing of Ensemble Forecasts, Elsevier, 49–89. https://doi.org/10.1016/B978-0-12-812372-0.00003-0

Wilks, D.S., 2019: Statistical Methods in the Atmospheric Sciences. 4th ed. Elsevier, Amsterdam.

Wind Europe, 2021: Wind energy in Europe 2020. Statistics and the outlook for 2021-2025. Available at: https://windeurope.org/intelligence-platform/product/wind-energy-in-europe-in-2020-trends-and-statistics/ [Accessed on 17 August 2021]

World Wind Energy Association, 2021: Worldwide wind capacity reaches 744 gigawatts – an unprecedented 93 gigawatts added in 2020. Available at: https://wwindea.org/worldwide-wind-capacity-reaches-744-gigawatts/ [Accessed on 17 August 2021]

Yuen, R. A., Baran, S., Fraley, C., Gneiting, T., Lerch, S., Scheuerer, M., and *Thorarinsdottir, T.L.,* 2018: R package ensembleMOS, Version 0.8.2: Ensemble Model Output Statistics. Available at: https://cran.r-project.org/package=ensembleMOS [Accessed on 17 August 2021]